Phylogenetic models to infer cancer evolution

The recent development of experimental technologies enables the study of vast data from the sequencing of the genetic material of individual cells. The cells inside a tumor can be divided into clones, that is, subpopulations with common mutation traits. Intratumor heterogeneity studies can be traced back to the XIX century when morphological differences between single tumor cells were observed under the microscope. Nowadays, we have the potential to infer information to provide a prognosis, explain mechanisms of drug resistance, and tailor cancer treatment.

To depict cancer evolution in a single patient we can infer the phylogenetic tree, in which leaves represent single cells, while internal nodes symbolize their hypothetical common ancestors. The single cell can be represented by a copy-number profile (CNP), that is, numbers of copies of reference chromosome parts. In our project, we assume that CNPs are given as input.

To compare the genetic material of cells, one needs a measure to describe how much two CNPs differ. One of approaches is the evolutionary distance, that represents the number of evolutionary events that transform one CNP into the other. It is not a distance from mathematical point of view, as it is not symmetric. In cancer evolution large parts of the genome can be multiplied or deleted. The most recent evolutionary distance, published in 2020, is modeling this phenomenon. However, computation of the distance is a difficult problem, that we call NP-hard in computer science. Previous simplified variant of evolutionary distance is easier to compute, however, the phylogeny inference problem under that simplified model is difficult (NP-hard). In summary, we deal with a difficult problem that consists of difficult subproblems, and there is a need for feasible models and solutions.

Our project will focus on two topics - to design and analyze:

- Phylogenetic models for data sampled from one patient in regular examinations,
- Models to discover regularity in evolutionary processes from multiple phylogenetic trees.

The first topic is especially innovative, as phylogenetic trees infer the past from currently available data. For example, to model the evolution we can analyze only the genomes of existing species, and paleontological data cannot provide the full picture of the past. Sampling from a patient in regular examinations will provide data of all cells in the population that exist at the time of the examination. Therefore, a new opportunity arises, to model the evolution between checkpoints. The process of evolution is so fast, that we can observe it.

Our theoretical analysis will employ techniques as standard mathematical proof techniques, graph theory, comparative genomics methods, phylogenetic inference. For all models, we will design and develop algorithms. Our research will include standard algorithmic design paradigms, run-time complexity analysis, and problem complexity study. We expect to face NP-hard problems. Therefore, we will employ techniques like hill-climbing heuristic, fixed-parameter tractability, computational approximation, and integer linear programming.